Exploiting CLIP Self-Consistency to Automate Image Augmentation for Safety Critical Scenarios

Sujan Sai Gannamaneni[1,3], Frederic Klein[2], Michael Mock[1], Maram Akila[1,3]
[1]Fraunhofer IAIS, [2]University of Bonn, [3]Lamarr Institute
{sujan.sai.gannamaneni,michael.mock,maram.akila}@iais.fraunhofer.de,
frederic.klein@uni-bonn.de

Abstract

With the current interest in deploying machine learning (ML) models in safety-critical applications like automated driving (AD), there is increased effort in developing sophisticated testing techniques for evaluating the models. One of the primary requirements for testing is the availability of test data, particularly test data that captures the long tail distributions of traffic events. As such data collection in the real world is hazardous, there is also a necessity for generating synthetic data using simulators or deep learning-based approaches. We propose a pipeline to generate augmented safety-critical scenes of the Cityscapes dataset using pre-trained SOTA latent diffusion models with additional conditioning using text and OpenPose-based ControlNet, where we have fine-grained control of the attributes of the generated pedestrians. In addition, we propose a filtering mechanism, similar to self-consistency checks in large language models (LLMs), to improve the quality of the generated data regarding the adherence to generated attributes, reaching $\sim 25\%$ *improvement in our experiments. Finally, using pre-trained SOTA segmentation models on Cityscapes, we evaluate the generated dataset's viability by qualitatively evaluating the predicted segmentation maps.*

1. Introduction

With the increased focus on AI applications in safety-critical systems like autonomous driving (AD), there is a growing need and interest in building trustworthy or Safe AI models. Upcoming standards like ISO/CD PAS 8800 [22] and projects like KI-Absicherung[1] propose considering functional insufficiencies where evidences from various testing approaches are used to build an overall safety argumentation for the AD vehicle [9]. One of the key ingredients for such an approach is the availability of assurance or test datasets [19, 38] which contain, in addition to normal AD scenarios, safety-critical scenarios or corner cases [6, 43] where the long tail distribution of driving scenarios is captured (*e.g*., a scene of a child running in front of a car). As such data acquisition is challenging and potentially dangerous to capture in the real world, there is an interest in generating synthetic or augmented datasets. Furthermore, the issues of data coverage and completeness are considered through operational design domains (ODDs) [25], which define the scope of AD vehicle use (see, *e.g*., [21] for a pedestrian focused ODD or Euro NCAP [13] for testing scenarios).

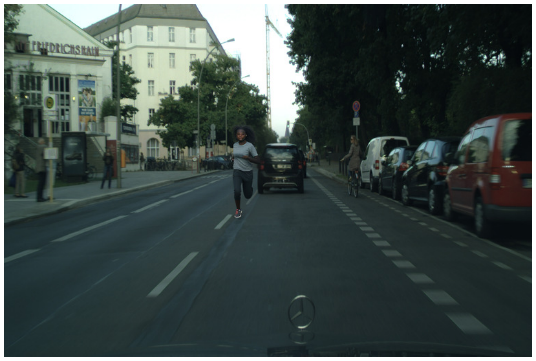


Figure 1. A sample image generated as part of our pipeline where an image from the Cityscapes [11] dataset is augmented with a pedestrian with attributes: *female, young, grey shirt-color, dark-skinned, running*.

While several recent works [12, 16, 17, 20, 28, 36, 38, 41] focus on using computer simulators to generate synthetic datasets for testing AD models, there is a gradual shift in recent years to using deep learning based approaches with the introduction of NeRFs [45] and diffusion models [7, 29]. While the former approaches allow for the easy generation of new data, creating maps and digital assets is cost-intensive, and the data still suffers from a lack of realism, leading to domain gap issues. The deep learning-based approaches not only present novel methods to generate crit-

[1] https://www.ki-absicherung-projekt.de/en/

ical scenarios in terms of new scenes but also augment existing datasets in critical ways, thereby staying close to the original data distribution leading to smaller overall domain gaps, compare also FID scores in [8]. However, as the generation of the safety-critical data requires not only realistic image generation capability but also granular control over the attributes of the objects inside the image, the deep learning approaches based on diffusion models like Stable Diffusion [34] can still suffer from quality issues. For example, if the goal is to generate a safety critical scene of *"a black woman walking and wearing a red shirt"* in front of the ego vehicle, incorrect generation of the different attributes of the person could lead to inconclusive tests.

We address this problem by designing a pipeline to augment images from existing datasets like Cityscapes with pedestrians to create novel safety-critical scenarios (*e.g*., see Fig. 1). The pipeline takes as input images from the Cityscapes dataset, textual prompts that contain the requested attributes of the pedestrian, and additional conditioning information like pedestrian pose. Using an approach similar to self-consistency techniques in large language models (LLMs) [40], we introduce a filtering mechanism that improves the quality of the generated data w.r.t. adherence to specification, *i.e*., images where the augmentation is correct w.r.t. the specification are more likely to remain in the final augmented dataset. We evaluate the quality of the augmented and filtered augmented datasets by human evaluation on a sampled subset. Our contributions can be summarized as follows:

- A pipeline for inpainting safety-critical pedestrians into real-world datasets like Cityscapes with control over granular attributes from the ODD.
- An automated self-consistency check over the augmented images to filter out most augmented images that do not adhere to required attributes and to also further understand the limitations of foundational models like CLIP.
- Qualitative evaluation of publicly trained semantic segmentation models on the augmented data to identify the viability of the dataset and also indicate potential weaknesses of some of the models.

2. Related Work

In this section, we discuss SOTA approaches in synthetic data generation and augmentation, both using computer simulators and DNN-based approaches with a focus on AD. Subsequently, we discuss how our proposed consistency check for CLIP [32] differs from self-consistency approaches in LLMs.

As the interest in data generation in AD for both training and testing has increased over the past years, several computer simulator-based synthetic datasets have been generated like VirtualKITTI [16], GTA5 [33], SYNTHIA [36], Valerie22 [20], Synscapes [41], and SynPeDS [38]. In addition, with the Carla [12] simulator, some works [17, 28] generated test datasets with additional metadata attributes and evaluated performance limiting factors of DNNs-under-test. While one can ensure granular control of the scene generation with both game engine and physical render-based approaches, the data generated can still suffer from low realism, leading to domain gap issues. Therefore, insights can be gained only about DNNs-under-test when they are both trained and tested exclusively on these datasets. To transfer any insights gained about the DNN-under-test to its behavior on real-world data, the field of domain adaptation and domain gap should be considered [39]. In addition, while these approaches make the generation of safety-critical scenarios more cost-effective than real-world data collection, the costs of creation of new digital assets and maps for both simulators and physical renderers limit the diversity of the generated data [45].

Due to these issues, there has also been an interest in generating synthetic data using deep learning-based approaches. While some approaches [23] proposed generating complete synthetic images with GANs, others proposed inpainting techniques for new object insertion or object modification in real-world logs or existing datasets to build new scenes with reduced domain gap. The Cityscapes dataset [11] has been augmented using GANs [31] and VAEs [2, 49]. Similarly, some approaches [1, 26] have used CAD model rendering and sensor simulation to design data generation pipelines for object insertion. More recently, neural radiance (NeRF)-based approaches like UniSim [45] allow for a granular manipulation of the entire scene by creating a digital twin of real-world logs. While the qualitative results look promising regarding realism, some of these approaches are not available for open access.

With the recent success of publicly available diffusion models, several works [7, 29, 42] use Stable Diffusion [34] models to generate data. While DatasetDM [42] generates images and perception annotations for multiple tasks, Metzen et al. [29] focus on generating images to identify systematic errors of pre-trained classification models w.r.t. ODDs. Boreiko et al. [7] is the closest to our approach where they propose a multi-stage pipeline for outpainting images where they have granular control of the different attributes of the objects. They argue that outpainting leads to improved image quality by removing hallucinations outside object boundaries. In contrast, we propose inpainting on existing real-world datasets like Cityscapes in our pipeline to improve the realism of augmented images while maintaining overall data distribution.

There are also several works for pose-guided person image synthesis, particularly for generating arbitrary poses with different attributes, with a focus on the fashion industry [4, 37, 44]. While [4, 37] both propose specialized architectures based on diffusion models for the arbitrary pose

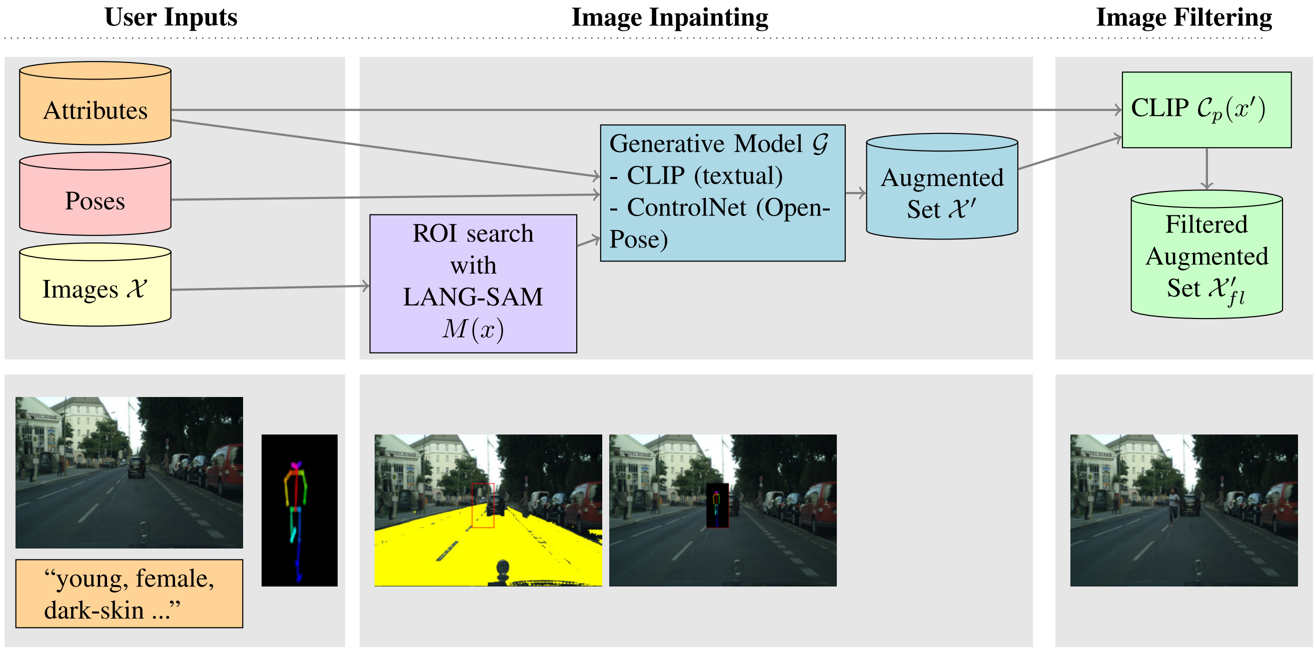

User Inputs
Image Inpainting
Image Filtering
Attributes
Poses
Images $\mathcal{X}$
ROI search with LANG-SAM $M(x)$
Generative Model $\mathcal{G}$
- CLIP (textual)
- ControlNet (OpenPose)
Augmented Set $\mathcal{X}'$
CLIP $\mathcal{C}_p(x')$
Filtered Augmented Set $\mathcal{X}'_{fl}$
"young, female, dark-skin ..."

Figure 2. Our proposed pipeline for generating safety-critical scenarios in AD datasets with self-consistency-based filtering to ensure a higher quality of the final augmented dataset. On the top, we provide an overview of the different stages of the pipeline. On the bottom, the inpainting pipeline is shown for a single image x, prompt t, and pose to obtain x'.

synthesis, Xu et al. [44] focuses on pose and attribute generation using GANs.

Self-consistency [40] has been proposed as a method to improve the performance of large language models by sampling multiple reasoning paths and then using some voting-based approach to choose the final prediction. Multi-modal foundational models like CLIP [32] can be used for different applications such as textual conditioning in Stable Diffusion [34] or zero-shot classification of attributes of objects in images [18]. Therefore, consistency-based approaches using CLIP can be used to improve the performance of the overall application (*e.g.*, improved quality of generated images from Stable Diffusion). Recent approaches [14, 15] proposed a similar idea of improving the performance of generation pipelines by CLIP-based checks. However, while these approaches filter based on the cosine similarity of the original and the generated image or generated image and class label, we extend the consistency check to evaluation at the attribute level to ensure that the generated augmented test images adhere to the specification.

3. Pipeline

In this section, we first provide a formal notation of the task and then describe the concrete pipeline we propose for augmenting real-world datasets with safety-critical scenes by inserting pedestrians using diffusion models and self-consistency.

3.1. Notation

Let $x \in \mathcal{X}$ be an image from a test set $\mathcal{X}$ which we intend to augment by inpainting safety-critical objects (*e.g.*, pedestrians) to construct an augmented test set $\mathcal{X}'$. For inpainting on an image x to generate its augmented counterpart x', we require both a region of interest (ROI) and the required attributes as part of the specifications of the safety-critical object. The choice of the ROI heavily influences the safety criticality of the x'. Either manually or by an automated process, the ROI is chosen for each x by obtaining the coordinates of the center and corresponding width and height of a mask $M(x)$. For the attributes, let $\mathcal{P}$ be the set of all semantic dimensions, *e.g.*, "gender" $\in \mathcal{P}$ where we call $\text{attr}_p(x')$ the *requested* attribute for x' w.r.t. to the dimension $p \in \mathcal{P}$, *e.g.*, $\text{attr}_{\text{gender}}(x') \in \{\text{male, female}\}$. Such a list of semantic dimensions and corresponding attributes for AD can be taken from the operational design domain (ODD) (*e.g.*, as proposed by Herrmann et al. [21]). Given a specification of requested attributes considering different dimensions (*e.g.*, *gender: male, shirt-color: red, age: old, skin-color: white*), we build a textual prompt t.

Given user inputs, we can use a conditional generative model $\mathcal{G}$ to generate an augmented image x'. Here, $\mathcal{G}$ can consist of multiple conditioning models to support the generative process (*e.g.*, text, pose, depth, . . .). Some of these models can, in addition to providing conditioning to $\mathcal{G}$, also be used for discriminate tasks like classification. For example, let model $\mathcal{C}(t)$ provide the textual conditioning to $\mathcal{G}$ so that the specification in terms of required attributes is produced in x' during image generation/inpainting.

However, there is no guarantee that the textual conditioning with t ensures that the generation process for x' follows the specification, *i.e*., violations in the form of out-of-subgroup (OOS) or out-of-class (OOC) [29] are possible. Therefore, while it is straightforward to generate a test set with the above mentioned steps, it is further necessary to filter $\mathcal{X}'$ to retain only the images where the specification is met (*i.e*., consistent) to maintain the quality. We refer to this filtered augmented set as $\mathcal{X}'_{fl}$. As human inspection is time intensive to evaluate such consistency, we propose to automate this process by using some classification model $\mathcal{C}_p(x')$ which classifies the attributes of dimension $\mathcal{P}$ in a given image. As we use the same model $\mathcal{C}$ for both textual conditioning of $\mathcal{G}$ and for attribute classification, we refer to our approach as being based on self-consistency.

To highlight the usefulness of self-consistency and to evaluate the classification model, we introduce the following terms w.r.t. x' and $\mathcal{P}$: (i) x' is *consistent true positive (ctp)* w.r.t. $\text{attr}_p(x')$ if $\text{attr}_p(x')$ is requested and the classification model $\mathcal{C}_p(x')$ classifies it as such, (ii) x' is a *consistent false negative (cfn)* w.r.t $\text{attr}_p(x')$ if $\text{attr}_p(x')$ is requested but the classification model $\mathcal{C}_p(x')$ does not classify it as such, and (iii) x' is a *consistent false positive (cfp)* w.r.t. $\text{attr}_p(x')$ if $\text{attr}_p(x')$ is not requested but the classification model $\mathcal{C}_p(x')$ classifies it as such. This allows us to quantify self-consistency using classification metrics, *i.e*., we can compute precision and recall. If the precision and recall were 100%, then there would be effectively no information gain using self-consistency. However, our evaluations (see Tab. 3) indicate that there are significant deviations in self-consistency which offer signals to perform filtering. For example, say *shirt-color: brown* is requested, *shirt-color: red* is generated in x' and *shirt-color: red* is classified by $\mathcal{C}_p(x')$. This would be an instance of consistent false negative and this signal can be used to filter x'. Concretely, the filtering restricts to the subset

$$\mathcal{X}'_{fl} = \{x' \,|\, x' \in \mathcal{X}' : \forall p \in \mathcal{P} : \text{attr}_p(x') = \mathcal{C}_p(x')\}\,,$$

which contains only those augmented images x' that are consistent true positives w.r.t. all attributes.

Considering the overall pipeline, two potential failure modes are possible: (i) errors in the diffusion process and (ii) errors in the classification model. To understand the quality of the filtered data and the filtering process with a focus on both these failure modes, we use human evaluation on a subset of $\mathcal{X}'$ to identify samples where the specification is met. From this, we can estimate the improvement in quality from the augmented dataset to the filtered augmented dataset.

3.2. Concrete Pipeline

Now, we describe the concrete pipeline and the different components required for the generation of the filtered augmented set $\mathcal{X}'_{fl}$ as shown in Fig. 2.

Region of Interest: We automate the search for the regions of interest $M(x)$ by using LANG-SAM,[2] a combination of the Segment Anything Model (SAM) [24] and Grounding-DINO [27] to segment the entire scene in x. Note that this step is not necessary if GT segmentation labels are available. After segmenting, we search for regions that are relatively "free", *i.e*., pixels belonging to classes *road* or *footpath*, and choose a random pixel from this selection. A rectangular mask $M(x)$ with this pixel as a center and pre-defined width and height is then chosen. We use multiple variants of width, height pairs while maintaining appropriate aspect ratio.

Object Inpainting with Diffusion models: Once the region of interest $M(x)$ is identified, for the generation of pedestrians with fine-grained attributes, we propose to use as generative model $\mathcal{G}$ a SOTA text-to-image latent diffusion model like Stable Diffusion [34]. As mentioned, $\mathcal{G}$ can be conditioned with various additional models. For inpainting, the input image x is encoded using an image encoder. The textual prompt containing the required attributes $\text{attr}_p(x')$ as part of the specification is encoded by text encoder of CLIP [32] *i.e*., model $\mathcal{C}(t)$, for conditioning the trained U-Net [35]. Furthermore, to ensure that the inpainting pedestrians have natural and realistic poses within the images and to have greater variety in poses, we use an additional conditioning model, *i.e*., ControlNet [46] with OpenPose [10]. As shown in Fig. 2, user provided pose images are overlaid in the ROI and ControlNet ensures that the inpainted image x' follows the provided input pose.

Filtering with Self-Consistency: We propose to filter the augmented dataset $\mathcal{X}'$ to improve the quality of the final filtered augmented set $\mathcal{X}'_{fl}$ by using self-consistency of CLIP [32], *i.e*., $\mathcal{C}_p(x')$. Earlier works have evaluated zero-shot classification capabilities of CLIP both at unseen datasets [32] and granular attributes [18]. Similar to the earlier work [18], we use prompt template of the form `['a photo of a {} person', 'a photo of a {} man', 'a photo of a {} woman', 'a photo of a {} guy', 'a photo of a {} lady']`, where the {} are replaced by either elements from `['young', 'younger']` or `['old', 'older']` to evaluate a semantic dimension, *e.g*., age using ensemble prompting [18, 32]. Given the original specification prompt t and its constituent attributes $\text{attr}_p(x')$ and the classifications of the CLIP model, we filter for the set of images that are consistent true positives (ctp) for all dimensions.

[2] https://github.com/luca-medeiros/lang-segment-anything/tree/main

4. Experiment Setup

We consider the test set of the Cityscapes dataset [11] to perform the proposed augmentation with safety-critical scenarios. The test set contains 1525 images with resolutions of 2048×1024 with semantic segmentation annotations for 19 classes relevant from an AD perspective. The images are resized to 768×512 before they are input to the LANG-SAM model for the ROI search.

For the proposed pipeline, as the inpainting diffusion model, we consider three publicly available custom checkpoints, namely *SD-v1.5*,[3] *Reliberate-v2*,[4] and *Deliberate-v5*.[5] For the ControlNet [46] component based on OpenPose [10], we choose the publicly available model *control-v11p-sd15-openpose*.[6] We make use of the code provided from Stable Diffusion web-ui [3] and adapt it as a python script for our pipeline. Regarding the hyper-parameters, we use the "Euler-a" sampler for the diffusion model and a classifier-free guidance scale (cfg) of 0.6. To foster reproducibility, the code, along with a detailed list of used prompts and hyper-parameters are provided.[7] We use a NVIDIA Tesla V100 for our experiments. On average, the pipeline takes ~ 8 seconds to inpaint a single image.

We also consider publicly available pre-trained models trained on Cityscapes dataset, such as two SETR [48] models, and one ICNet [47] model, as DNNs-under-test and perform inference on the filtered augmented set $\mathcal{X}'_{fl}$. We use the inference pipeline and the model weights from the mmsegmentation repository [30]. For the first SETR model, we consider a model with relatively high performance with a mIoU of 79.21 on the unaugmented Cityscapes test set trained with a resolution of 768×768 and a batch size of 8. The backbone is a ViT-L with training method SETR PUP [48]. As a contrast, we choose a relatively weaker SETR model that has been trained on fewer, *i.e.*, 16k, iterations and has a mIoU of 60.00. To contrast with the more recent SETR architecture, we choose a slightly "older" ICNet model, which has a mIoU of 68.14 trained with images of resolution 832×832 and uses a ResNet-18 backbone.

5. Results

In this section, we evaluate the three Stable Diffusion checkpoints, identify a suitable one, and then evaluate the quality of the generated augmented data from our pipeline and the improvements due to self-consistency. Subsequently, we evaluate the three different DNNs-under-test on the generated augmented images and discuss the viability of the generated data and the observed DNN performances.

	Size	% of Consistent Images (Eval. by Humans)
Augmented Set $\mathcal{X}'$	1439	69.23%
Filtered Augmented Set $\mathcal{X}'_{fl}$	649	87.5%

Table 1. Human evaluation of 10% of the augmented set and the filtered augmented set highlighting the benefits of the self-consistency based filtering approach.

5.1. Comparison of different Stable Diffusion checkpoints

The inpainting quality of our pipeline is heavily influenced by the choice of the generative model $\mathcal{G}$. Among the publicly available latent diffusion models, Stable Diffusion [34] is a popular choice [7, 29]. However, as variants in terms of model weights are available for Stable Diffusion, the choice of the variant could also have an influence on the quality of the final output. Therefore, using weights from three different checkpoints of Stable Diffusion, *i.e.*, *SD-v1.5*, *Reliberate-v2*, and *Deliberate-v5*, we perform a qualitative evaluation of the generated data to identify a suitable model. To generate the data, we augment 1525 test images from the Cityscapes dataset. Here, as part of the specification for requested attributes $\text{attr}_p(x')$, we consider five semantic dimensions (*i.e.*, *age*, *gender*, *skin-color*, *shirt-color*, *action*), which we reflect in the generated prompt t as "`a {age} {gender} {skincolor} person:1.5 wearing {shirtcolor} colored shirt {action}, full body shot`" (similar to other works [7, 29]). The values per dimension are shown in Tab. 3 and, in combination, result in 112 different prompts.

Fig. 3 shows three samples each from the three checkpoints. In terms of the inpainting, based on random sampling of the generated data, we notice that it is more likely that inpainting fails with the *SD-v1.5* checkpoint (see first two rows for *SD-v1.5* in Fig. 3). Furthermore, the inpainting quality in comparison to the other two models appears to be "poor" in terms of realism (see also Appendix A). As human evaluation of the samples yields no strong qualitative difference between the two latter checkpoints, we select the more recent *Deliberate-v5* model.

5.2. Evaluation of the pipeline

In this experiment, we evaluate the improvement in quality achieved by our filtering mechanism using self-consistency when applied to the augmented set $\mathcal{X}'$ generated using the *Deliberate-v5* checkpoint. Based on the consistency check, we filter for images that are consistent true positives (ctp) for all attributes using CLIP to obtain the filtered augmented

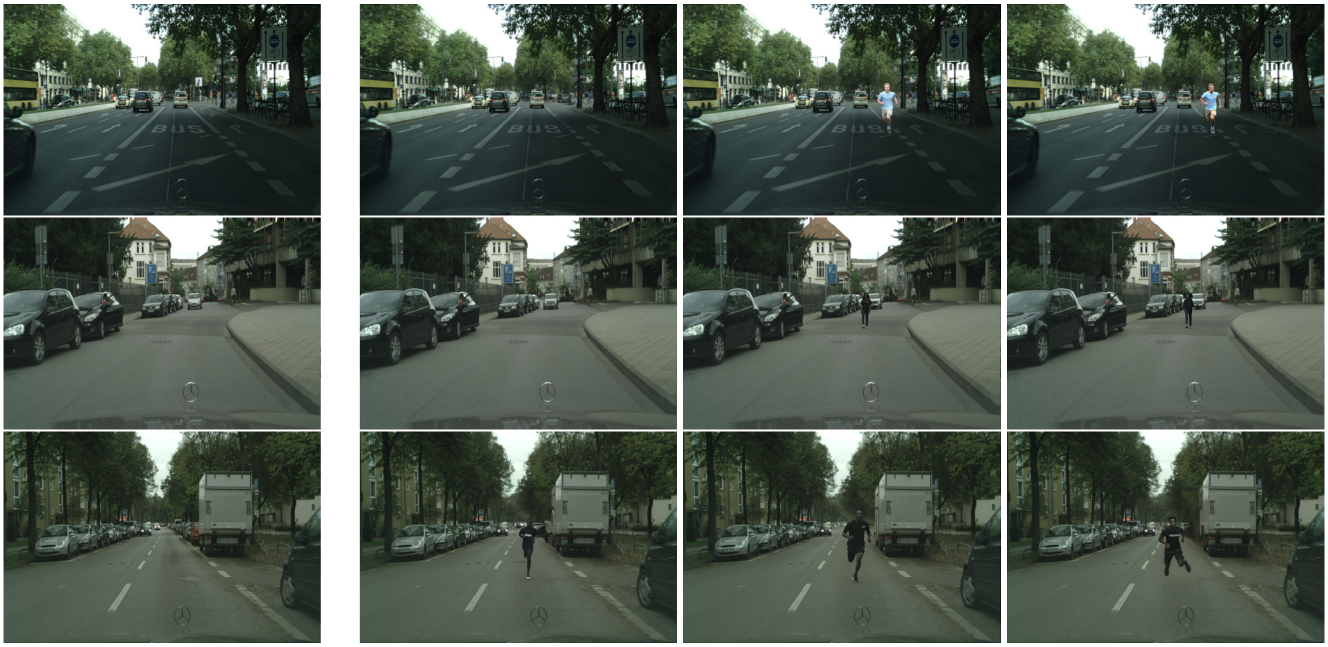
[3] https://huggingface.co/runwayml/stable-diffusion-v1-5

[4] https://huggingface.co/XpucT/Reliberate

[5] https://huggingface.co/XpucT/Deliberate

[6] https://huggingface.co/lllyasviel/ControlNet-v1-1/tree/main

[7] https://github.com/sujan-sai-g/Exploiting-CLIP-Self-Consistency-to-Automate-Image-Augmentation-for-Safety-Critical-Scenarios



(a) Input Image (b) *SD-v1.5* (c) *Reliberate-v2* (d) *Deliberate-v5*

Figure 3. Samples of the augmented images with inpainted pedestrians using three different checkpoints of latent diffusion models [34]. Note that in some instances, *e.g.*, the top 2 images in *SD-v1.5*, inpainting could completely fail.

	Precision	Recall	F1-Score	Support
Inconsistent Images	0.47	0.84	0.61	44
Consistent Images	0.89	0.59	0.71	99

Table 2. Using human evaluation as ground-truth information, we test the efficacy of the filtering approach by framing it as a classification problem. 10% of samples from augmented $\mathcal{X}'$ that are evaluated by humans is compared against the self-consistency results.

set $\mathcal{X}'_{fl}$ and the size of both sets is shown in Tab. 1.[8] Nearly half of the images were filtered either because of inconsistent attributes or because no pedestrian was actually generated at all. To ensure that there is, in fact, an improvement in the quality, we perform a human evaluation of 10% of $\mathcal{X}'$ and $\mathcal{X}'_{fl}$, respectively. The human evaluator checks if the augmented image contains a pedestrian with the specified attributes, *i.e.*, now the consistency is checked by a human. We see a significant improvement of about $\sim 25\%$ in quality using our self-consistency filtering approach.

Furthermore, we evaluate the quality of the filtering process itself, as discussed in Tab. 2. Here, the 143 (10% of 1439) samples from $\mathcal{X}'$, which are evaluated by humans (ground-truth) are compared against the self-consistency results of CLIP (predictions). Therefore, the precision shown in Tab. 2 should correspond to the % of consistent images in $\mathcal{X}'_{fl}$ shown in Tab. 1. However, there is a slight deviation ($\sim 2\%$) between both the values as the human evaluation is done on independent sets. While the filtering process through self-consistency is good at detecting inconsistent images with a high recall of 84%, a lot of consistent images are filtered out. However, as data generation is relatively cheap, higher recall on inconsistent images is of higher importance to obtain a higher quality $\mathcal{X}'_{fl}$.

We further evaluate the self-consistency of CLIP by calculating the ctp, cfp, and cfn (as defined in Sec. 3) for each of the attributes before we perform the filtering. Based on these values, we calculate the precision, recall, and f1-score for self-consistency in Tab. 3 on the augmented set $\mathcal{X}'$. Consider the dimensions *age* and *gender*. As the precision and recall values are relatively high for the attributes within these dimensions, we can conclude that the textual conditioning and model classification are highly consistent for these dimensions, implying that the classification and generation processes agree. This could suggest that both CLIP and diffusion models are good at classifying and generating common attributes. In contrast, consider the dimension *shirt-color* where significant deviations in the recall can be seen for certain attributes like *brown* and *grey*. In Fig. 4, we show examples of inconsistencies for shirt-colors, which show that the textual conditioning fails in cer-

[8] There is a deviation between the test set size of Cityscapes (1525) and the augmented set $\mathcal{X}'$ (1439) as some images are skipped during the ROI search process as no mask $M(x)$ can be found.

Sem. Dimension		Support	Precision$_{SC}$	Recall$_{SC}$	F1-Score$_{SC}$
Age	Young	732	0.89	0.84	0.86
	Adult	707	0.84	0.89	0.86
Gender	Female	717	0.90	0.90	0.90
	Male	722	0.90	0.90	0.90
Skin-colour	Dark-skinned	715	0.77	0.93	0.84
	White-skinned	724	0.91	0.72	0.80
Shirt-colour	Red	188	0.91	0.79	0.84
	Blue	197	0.75	0.80	0.78
	Green	231	0.92	0.85	0.88
	Black	201	0.56	0.59	0.57
	White	207	0.78	0.60	0.68
	Brown	195	0.69	0.05	0.09
	Grey	220	0.36	0.76	0.49
Action	Walking	734	0.80	0.94	0.87
	Running	705	0.93	0.76	0.83

Table 3. For the generated dataset before filtering, we estimate the consistency at attribute level using the defined consistent true positive (ctp), consistent false positive (cfp), and consistent false negative (cfn) values.

tain cases where the shoe or pants are red instead of the shirt. Similar failures can also occur when the classification itself fails, which was also discussed in an earlier work [18].

5.3. Insights on pre-trained models

In this experiment, we evaluate the viability of $\mathcal{X}'_{fl}$ by performing inference on this data using publicly available pre-trained models. As we do not generate the corresponding ground-truth for $\mathcal{X}'_{fl}$, we perform a qualitative evaluation of the segmentation outputs. To study the viability of $\mathcal{X}'_{fl}$, we can consider the following two requirements: (i) the generated images need to be close to the original data distribution to reduce issues related to domain gap, and (ii) the generated test set should be challenging for the DNNs-under-test to gain some insights about their failures. To test this, we choose two SETR models, one with very high performance and one with relatively lower performance on the standard Cityscapes test set. In addition, we choose an ICNet model, which is a relatively older architecture. We deliberately construct augmented images that are relatively easy to detect, *i.e.*, the pedestrians are in front of the ego-vehicle without any occlusions. As can be seen from Fig. 5 and Appendix B, while the stronger SETR model segments the inpainted pedestrians without any issues, the weaker model and the older architecture have artifacts in the predictions for these "easy" pedestrians, while the overall segmentation for other classes remains reasonable. This shows the potential viability of the dataset for uncovering the weaknesses of the latter models.

6. Conclusion

We introduced a pipeline for the generation of safety-critical scenarios in AD datasets using inpainting with diffusion models conditioned with text and pose. The pipeline offers fine-grained control over the generation of different attributes of safety-critical pedestrians, enabling the testing for the systematic failure modes of different DNNs-under-test. In addition to the augmented images, this method also enables the generation of granular metadata about the inpainted objects, which can be used for downstream evaluation tasks like identifying systematic weaknesses [18]. Furthermore, we show that the performance of the pipeline can be significantly improved by employing a self-consistency check using CLIP to ensure that the inpainted images contain the requested attributes. In our experimental results, we first perform a qualitative evaluation to find the best checkpoint for our proposed task. Then, we evaluate the impact of the proposed self-consistency check by human evaluation of a subset of the unfiltered and filtered data. Our evaluation shows that self-consistency leads to $\sim 25\%$ improvement in the quality of filtered data. We further show the viability of the generated data by qualitative evaluation of the inference results of pre-trained segmentation models.

Limitations and Future work: While diffusion models show significant capabilities, certain semantic dimensions like occlusion still prove challenging. Modifying the pipeline to enable multi-step inpainting might solve this issue, but we leave this for future work. Similarly, the pipeline can be extended to other road users, *e.g.*, using different poses and prompts for wheel-chair users or edge-based conditioning to capture bikes. Although there has been an improvement in photo realism, certain artifacts can

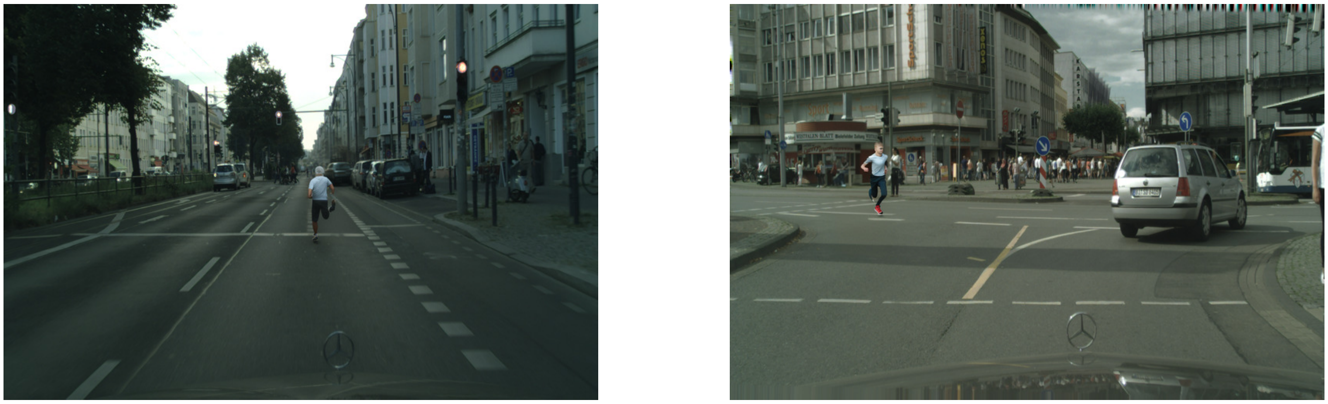


Figure 4. Samples of inconsistent images which we aim to filter from the augmented set. Left: specification *shirt-color* is brown but generated color is white, Right: specification *shirt-color* is red but generated color is white.

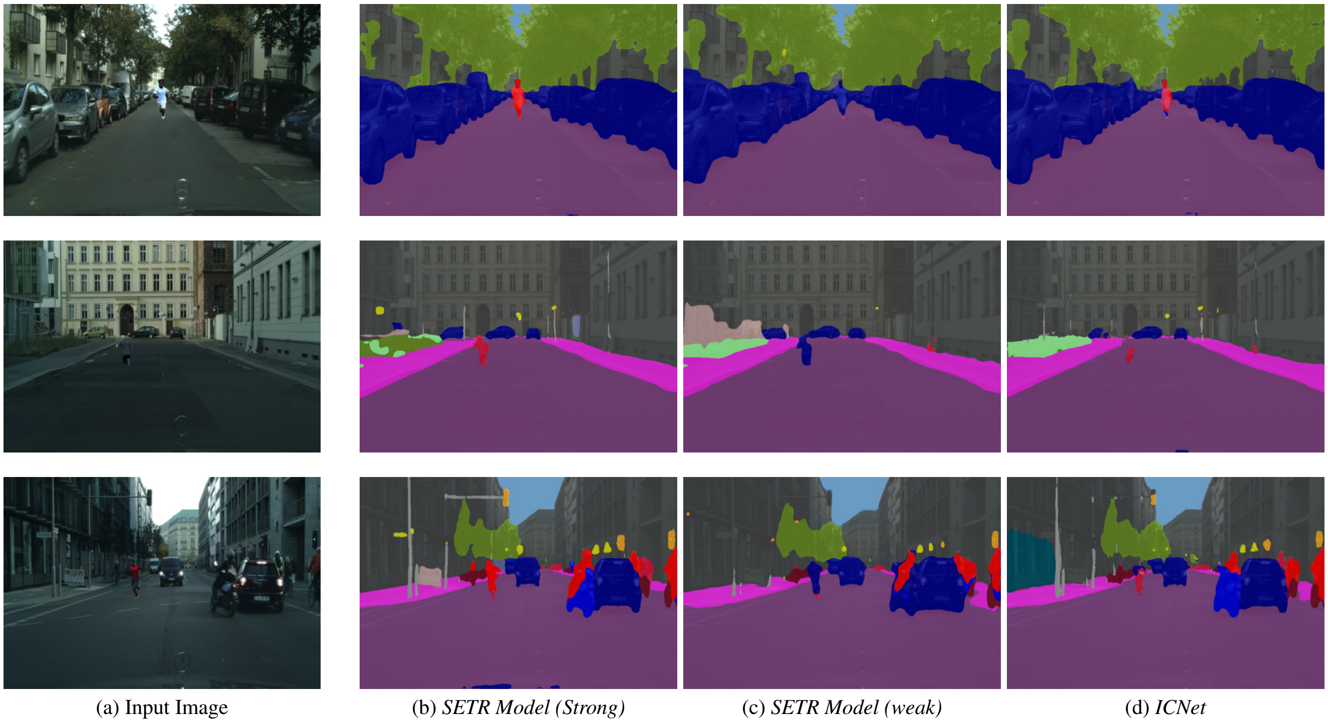

(a) Input Image
(b) *SETR Model (Strong)*
(c) *SETR Model (weak)*
(d) *ICNet*

Figure 5. Performance of three pre-trained models on images from the filtered augmented set.

still occur during the generation of face or body parts. Similarly, for certain colors, like red or green, the image saturation in the inpainting area does not fit the rest of the image. With improvement in diffusion model quality, such issues can be mitigated. Concerning fairness and bias, the given specification could introduce design- or selection-biases. Furthermore, the produced images can sometimes carry societal biases [5]. When this pipeline is used in an industrial context, care must be taken to ensure such biases do not transfer into the generated datasets. For a quantitative evaluation of DNNs-under-test, corresponding GT data is also required. While it is possible to generate bounding boxes based on the ROI, detailed segmentation masks might be possible based on the input pose which we intend to explore in future work.

7. Acknowledgments

This work has been funded by the European Union and the German Federal Ministry for Economic Affairs and Climate Action as part of the safe.trAIn project. The authors would like to thank Daniel Becker for fruitful discussions on the technical setup.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126:961–972, 2018. 2

[2] Pierfrancesco Ardino, Yahui Liu, Elisa Ricci, Bruno Lepri, and Marco De Nadai. Semantic-guided inpainting network for complex urban scenes manipulation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9280–9287. IEEE, 2021. 2

[3] AUTOMATIC1111. Stable Diffusion Web UI, 2022. 5

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 2

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. 8

[6] Jan-Aike Bolte, Andreas Bar, Daniel Lipinski, and Tim Fingscheidt. Towards corner case detection for autonomous driving. In *2019 IEEE Intelligent vehicles symposium (IV)*, pages 438–445. IEEE, 2019. 1

[7] Valentyn Boreiko, Matthias Hein, and Jan Hendrik Metzen. Identifying systematic errors in object detectors with the scrod pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4090–4099, 2023. 1, 2, 5

[8] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2, 2023. 2

[9] Simon Burton, Christian Hellert, Fabian Hüger, Michael Mock, and Andreas Rohatschek. *Safety Assurance of Machine Learning for Perception Functions*, pages 335–358. Springer International Publishing, Cham, 2022. 1

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2021. 4, 5

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 1, 2, 5

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 1, 2

[13] Euro NCAP. European new car assessment programme AEB/LSS VRU systems. https://www.euroncap.com/media/80156/euro-ncap-aeb-lss-vru-test-protocol-v451.pdf, 2024. Accessed: 2024-04. 1

[14] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1257–1266, 2024. 3

[15] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 3

[16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[17] Sujan Gannamaneni, Sebastian Houben, and Maram Akila. Semantic concept testing in autonomous driving by extraction of object-level annotations from carla. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1006–1014, 2021. 1, 2

[18] Sujan Sai Gannamaneni, Arwin Sadaghiani, Rohil Prakash Rao, Michael Mock, and Maram Akila. Investigating clip performance for meta-data generation in ad datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3840–3850, 2023. 3, 4, 7

[19] Christoph Gladisch, Christian Heinzemann, Martin Herrmann, and Matthias Woehrle. Leveraging combinatorial testing for safety-critical computer vision datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 324–325, 2020. 1

[20] Oliver Grau and Korbinian Hagn. Valerie22-a photorealistic, richly metadata annotated dataset of urban environments. In *Proceedings of the 7th ACM Computer Science in Cars Symposium*, pages 1–9, 2023. 1, 2

[21] Martin Herrmann, Christian Witt, Laureen Lake, Stefani Guneshka, Christian Heinzemann, Frank Bonarens, Patrick Feifel, and Simon Funke. Using ontologies for dataset engineering in automotive ai applications. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 526–531. IEEE, 2022. 1, 3

[22] ISO/CD PAS 8800:2023-10. Road vehicles - safety and artificial intelligence. Standard, International Standards Organisation (ISO), Geneva, CH, 2023. Under development. 1

[23] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5820–5829, 2021. 2

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and

Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 4

[25] Philip Koopman and Frank Fratrik. How many operational design domains, objects, and events? In *SafeAI@AAAI*, 2019. 1

[26] W. Li, C. W. Pan, R. Zhang, J. P. Ren, Y. X. Ma, J. Fang, F. L. Yan, Q. C. Geng, X. Y. Huang, H. J. Gong, W. W. Xu, G. P. Wang, D. Manocha, and R. G. Yang. AADS: Augmented autonomous driving simulation using data-driven algorithms. *Science Robotics*, 4(28):eaaw0863, 2019. 2

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[28] Maria Lyssenko, Christoph Gladisch, Christian Heinzemann, Matthias Woehrle, and Rudolph Triebel. Instance segmentation in carla: Methodology and analysis for pedestrian-oriented synthetic data generation in crowded scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 988–996, 2021. 1, 2

[29] Jan Hendrik Metzen, Robin Hutmacher, N. Grace Hua, Valentyn Boreiko, and Dan Zhang. Identification of systematic errors of image classifiers on rare subgroups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5064–5073, 2023. 1, 2, 4, 5

[30] MMSegmentation Contributors. OpenMMLab Semantic Segmentation Toolbox and Benchmark, 2020. 5

[31] Xi Ouyang, Yu Cheng, Yifan Jiang, Chun-Liang Li, and Pan Zhou. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. *arXiv preprint arXiv:1804.02047*, 2018. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[33] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 5, 6, 13

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 4

[36] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[37] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*, 2023. 2

[38] Thomas Stauner, Frederik Blank, Michael Fürst, Johannes Günther, Korbinian Hagn, Philipp Heidenreich, Markus Huber, Bastian Knerr, Thomas Schulik, and Karl-Ferdinand Leiß. SynPeDS: A synthetic dataset for pedestrian detection in urban traffic scenes. In *Proceedings of the 6th ACM Computer Science in Cars Symposium*, pages 1–10, 2022. 1, 2

[39] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 2

[40] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[41] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 1, 2

[42] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[43] Zhang Xinxin, Li Fei, and Wu Xiangbin. CSG: Critical scenario generation from real traffic accidents. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1330–1336. IEEE, 2020. 1

[44] Cheng Xu, Zejun Chen, Jiajie Mai, Xuemiao Xu, and Shengfeng He. Pose-and attribute-consistent person image synthesis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–21, 2023. 2, 3

[45] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023. 1, 2

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4, 5

[47] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[48] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 6881–6890, 2021. 5

[49] Rong Zhi, Zijie Guo, Wuqiang Zhang, Baofeng Wang, Vitali Kaiser, Julian Wiederer, and Fabian B Flohr. Pose-guided person image synthesis for data augmentation in pedestrian detection. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1493–1500. IEEE, 2021. 2